Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 57 (2015) 1124 – 1130

Procedia Computer Science

3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015)

An Advanced Multi Class Instance Selection based Support Vector Machine for Text Classification

Ramesh B[a*], Dr.J.G.R.Sathiaseelan[b]

[a]*Department of Computer Science, Bishop Heber College, Tiruchirapalli, TN, India. ram.7110@gmail.com*
[b]*Head, Department of Computer Science, Bishop Heber College, Tiruchirapalli, TN, India. jgrsathiaseelan@gmail.com*

Abstract

Machine learning techniques is most commonly used technique in text mining. Support Vector Machine (SVM) is a most useful supervised learning technique for text classification. In this paper we proposed advanced Multi Class Instance Selection based support vector machine (AMCISSVM) to increasing efficiency of support vector machine. The proposed algorithm is compared with Multi Class Instance Selection (MCIS) and Neighborhood Property based Pattern Selection (NPPS) algorithm. The advanced MCIS has shown high accuracy to multi datasets. These experimental datasets are retrieved from UCI machine learning repositories.

Peer-review under responsibility of organizing committee of the 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015)

Keywords: Text Mining, Classification, Support Vector Machine, Instance Selection;

1. Introduction

Nowadays, the growth of IT has caused the usage of electronic text documents huge. Due to this reason, text mining is very useful technique to retrieve the interesting knowledge from large volume of text documents. Extracting of relevant knowledge in huge text documents is a difficult one. Extracting accurate knowledge in text documents is to facilitate users what they required.

* Corresponding author. Tel.: +91-97- 91-644212.
E-mail address: ram.73110@gmail.com.

Peer-review under responsibility of organizing committee of the 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015)
doi:10.1016/j.procs.2015.07.400

Text mining is successfully applied to several areas such as market analysis, business management, e-newspaper, web page classification and so on. Ning Zhong et.al.[1] proposed innovative and effective methods for Information Retrieval. It is provided different term-based methods to overcome this issue.

A document is a group of text collections. Text categories may be derived from a sparse classification scheme or from a large collection of very specific content identifiers. Text categories may be expressed numerically or as phrases and individual words. Traditionally, text categorization task is performed manually by domain experts. Also, text categorization is more time consumption. Last few years, text classification approach classified into two ways: one is knowledge engineering and another is machine learning. Currently, machine learning approach is sophisticated method in text classification which is useful to several domains.

Text classification is most popular part in text mining which means automatically text documents can be classified into inter related categories. The procedure of text classification is to build a classifier based on some labeld documents and to organise unlabeled documents into the pre-specified categories. At present, different techniques are used for text classification such as decision tree, naive bayes method and so on. However, Support Vector Machine (SVM) is trendy research topic in supervised learning classifier specially for text classification. Joao ventura et al.[2] proposed Concept-Extractor for multi word relevant expression. Relevant expressions are applied in many fields namely information retrieval and text document classification, text document clustering. Text classification methods are used to identify client sentiment and opinions to social networks and e-commerce sites. García et al.[3] Conducted an experimental study uses different machine learning techniques namely naive bayes, KNN, SVM and Rocchio classification. As a final point, they concluded that support vector machine classification accuracy is better than others. In this study, different criteria are followed to evaluate the text performance such as speed, scalability, accuracy, time complexity and flexibility.

Jiali et al.[4] introduced multi layer text classification framework with two level representation model for text classification. They proposed a context-based method for identifying most relevant concept to each term based on the document structure information. Gary Doran et al. [5] analyzed a detailed study of support vector machine for multiple instance classification with kernel methods and elaborately discussed sound and completeness of it. Selcuk korkmaz et al. [6] conducted an experiment support vector machine with different pre-processing methods namely single SVM, SVM-Recursive Feature Elimination, Wrapper Method and Subset Selection. However, this experiment conducted with small sample size and its training set is quite balanced.

The rest of this paper is organized as follows: Section II literature survey, Section III proposed work of this paper, Section IV experimental results and some conclusions with future work are drawn in Section V.

2. Literature Review

Support Vector Machine is one of the pioneered supervised learning classification techniques, which is highly applied to several areas such as hand written digit recognition, bio informatics, face recognition and text categorization. The vast growth of IT also increases the size of text document repository. Hence, the huge and the storage space and computational cost of model learning is very high. To overcome this problem, instance selection is one solution to solve these limitations. The classical SVM is heuristically extended to multiple instance classification in different dimensions.

Some instance selection algorithms are based on clustering technique. Subhransu Maji et al. [7] proposed additive kernel SVMs for efficient image classification. This additive kernel SVM is provides a better accuracy when compared to state of art classification techniques. Jingnian Chen et al. [8] offered Multi Class Instance Selection (MCIS) method to choose instances nearest boundary and MCIS has been used to increase the speed of support vector machine. On the other hand, using this method selection of instances will be minimum.

Chih-Fong Tsai et al. [9] introduced a novel instance selection method for pre-processing namely Support Vector Oriented Instance Selection (SVOIS) for instance selection text classification. SVOIS compared with several instance selection algorithms such as ENN, IB3, ICF and DROP3. In future, this method is enhanced to non-linear regression plane and large scale experiments. Ramesh B et.al. [10] suggested pre-computed multi class instance selection(PCMCIS) for text classification to speed up SVM. Xinjun Peng et al. [11] proposed structural regularized projection twin support vector machine, which is extended to the nonlinear case by the kernel trick and woodbury

matrix identity. This method is well suited for binary classification. Qiaolin Ye et al. [12] developed multi weight vector support vector machine in linear case. But, MWVSVM is not extended to non-linear case and its single projection weight vector is not enough for each class to achieve best classification. Chin Heng Wang et al. [13] proposed hybrid classification approach using K-Nearest Neighbor and Support Vector Machine (SVM-NN). In this SVM-NN approach is consuming more time and less accuracy. Marcin et al.[14] designed multi test decision tree approach for micro array data classification. MTDTs are reasonably easy to examine and much more calculating in modeling heavy dimensional microarray data. This algorithm cause an increase in value of decision trees on gene expression data. Chih-Fong Tsai et al. [5] proposed Biological Genetic Algorithm (BGA) for instance selection of text classification. BGA compared with other text classification algorithms namely IB3, DROP3 and ICF. BGA consuming least computational time for instance selection. However, increasing the performance of BGA to huge datasets to develop a more efficient fitness function.

3. Proposed Work

In this paper, we present AMCISSVM to select instance from multi class datasets for increasing speed and efficiency of SVM. NPPS is used to select boundary instances using neighbour entropy concept. NPPS performance is fully based on the number instances selected. MCIS is used to clustering technique to select instances. Advanced MCIS has been used improved kernel to measure the distance between the origin and instances. Then an advanced MCIS has been tested with standard support vector machine formulation as using equ-1. Given a training set of instance-label pairs (x_i, y_i); i = 1……l where xi $\in R^n$ and y $\in \{1,-1\}^l$, the support vector machines (SVM) require the solution of the following standard support vector machine optimization formulation in equation [1]:

$$\min_{\omega, b, \varepsilon,} \frac{1}{2} y\| \omega\|^2 + C \sum_i \varepsilon_i, \qquad (1)$$

Show that $y_i((\omega, \phi(x_i))+b) \geq 1-\varepsilon_i$,

$\varepsilon_{i \geq 0}$.

Training vectors x_i are mapped into a higher dimensional space by the function ϕ. SVM finds a linear separating hyper plane with the maximal margin in this higher dimensional space. C > 0 is the penalty parameter of the error term. Furthermore, $K(x_i, x_j) \equiv (x_i)^T(x_j)$ is called the kernel function.

Classification accuracy is calculated as using equation (2).

$$\text{Accuracy} = \frac{\text{Number of selected instances}}{\text{total number of instances}} \text{ X } 100 \qquad (2)$$

To evaluate the performance of proposed AMCISSVM we compared it with another two instance selection algorithms namely NPPS and MCIS. These algorithms have been used for increasing performance of support vector machine. The flow of proposed AMCISSVM is explained in below algorithm. The following aspects are evaluated in or experiment.

1. Classification Accuracy
2. Ratio of selected instances
3. Efficiency of each algorithm.

AMCISSVM Pseudo code

1. For every class c, $1 \leq c \leq l$
2. $S_c \leftarrow \{x | x \text{ is from class } c\}$
3. Perform k-means cclustering on class c and get clusters Origin $O_1, O_2, \ldots O_n$
4. For every center O_i
5. $K(x_i, x_j) \equiv (x_i)^T(x_j)$

6. Compute $d(O_i,x)$ between O_i and every instance x;
7. If $N_c \geq r.N/3+n$
8. Get n_c instances in s_c closest to O_i and delete them from s_c, where $n_c = (N_c\text{-}r.N/3)/n$
9. End if
10. For every class $l \neq c$, $1 \leq l \leq L$
11. Search n_l instances of class l with least distance metrics and select them into s_c, where $n_l=(r \,.\, N - n_c)/(k.(L-1))$ and

$$n_c=\begin{cases} r.\frac{N}{3}, & Nc \geq r.N/3+n \\ Nc, & \text{otherwise} \end{cases} \quad (1)$$

4. Experimental Study

All experiments were performed on an Intel(R) Pentium(R) CPU P6200 running @2.13 GHz and 2GB RAM. We were used MATLAB version 7.6 in windows 7 operating system to all experiments.

4.1 Experimental Datasets

The AMCISSVM performance is experimentally proved using several medical disorder datasets retrieved from UCI repositories. The results of the AMCISSVM is compared with multi class instance selection (MCIS) and neighborhood property based pattern selection (NPPS) algorithm. For all data set, contains five aspects are listed: size of the data set, number of classes, number features, training set and testing set. The AMCISSVM is increasing performance to text classification. AMCISSVM was investigated using publicly available micro array datasets are described in Table I. These datasets are received from UCI machine learning repository.

Table I Experimental Datasets

Dataset	Class	Size	Features	Training Set	Test Set
Glass	6	214	9	172	42
Diabetes	2	768	8	603	165
Ionosphere	2	351	34	281	70

4.2 Experimental Setup

On each dataset, the setting of support vector multi class classification with kernel degree is 2, sigma is 2e-4, coefficient of kernel is 0 and cost parameter is 2.

4.3 Analysis of Results

4.3.1 Classification Accuracy

Below TABLE II shows AMCISSVM performance is better than MCIS and NPPS to all datasets. AMCISSVM is increasing classification accuracy to medical disordder datasets. Due to comparing classification accuracy AMCISSVM is provided better soluttion. The AMCISSVM performance is more in classification of micro array medical disorder data sets. Fig.1 shows that the variation of classification accuracy of NPPS, MCIS and AMCISSVM.

Table II Classification Accuracy

Data set	NPPS	MCIS	AMCISSVM
Glass	58.70	65.22	67.1224
Diabetes	76.13	78.71	78.9042
Ionosphere	91.67	94.44	94.7869

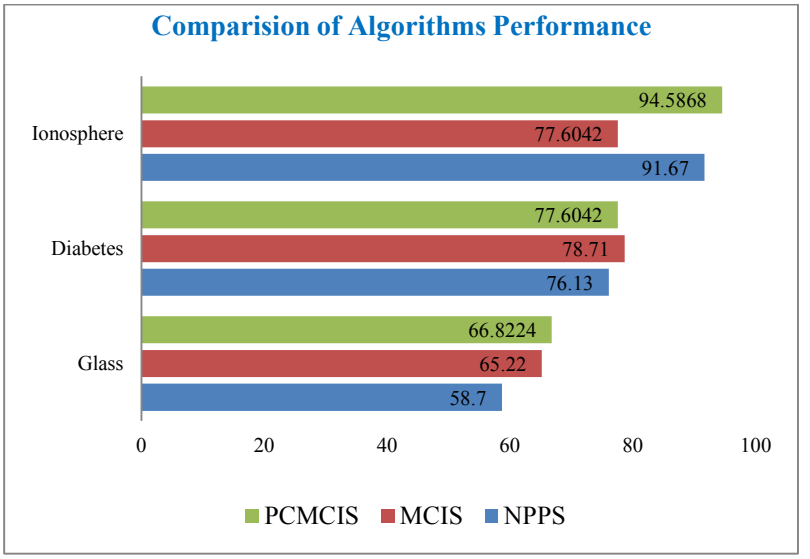

Fig.1. Comparison of Classification Accuracy

4.3.2 Comparison of ratio of selected instances

Below TABLE III shows the ratio of selected instances using NPPS, MCIS and AMCISSVM. AMCISSVM is selected more relevent instances. AMCISSVM performance is best compare to MCIS and NPPS in selected instances. AMCISSVM is indentifying more suitable optimum hyper plane with less time. Fig.2 shows the comparison of ratio of seleted instances of NPPS, MCIS and AMCISSVM.

Table III Ratio of Selected Instances

Data set	Glass	Ionoshere	Diabetes
NPPS	0.62	0.24	0.63
MCIS	0.52	0.55	0.65
AMCISSVM	0.67	0.95	0.65

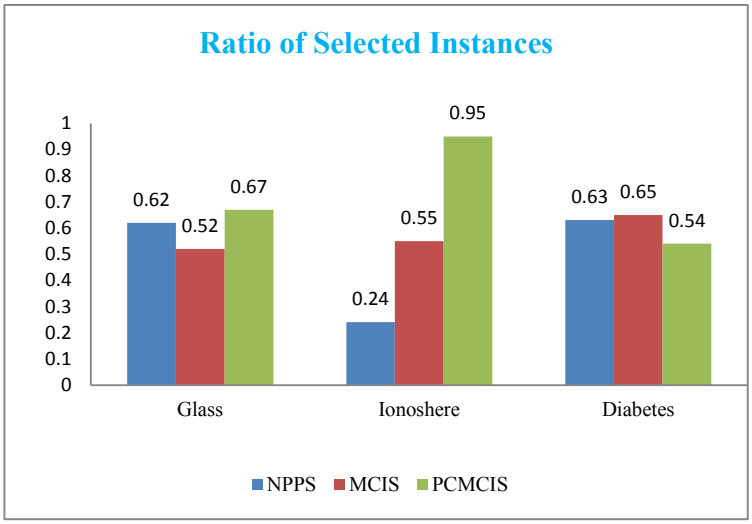

Fig.2. Comparison of ratio of selected instances.

5. Conclusion

Text mining techniques is more widely used to organize the high volume of literature by extracting exciting useful information. In this paper, we introduced advanced multi class instance selection namely advanced multi class instance selection based support vector machine. According to observation of experimental results AMCISSVM is better than NPPS and MCIS for selecting instances to classification . The AMCISSVM is best in all aspects such as classification accuracy, ratio of selected instances and time consumption. The proposed AMCISSVM is increasing accuracy to support vector machine. In future, we are increasing AMCISSVM accuracy with very less time and its performance is investigated with huge and different domain datasets.

References

1. Ning Zhonng, Yuefeng Li, and Sheng –Tang Wu, Effective Pattern Disvovery for Text Mining, IEEE Transactions on Knowledge and Data Engneering, Volume 24, No. 1, (2012).
2. Joao Ventura, Joaquim Silva, Mining Concepts from Texts, Procedia Computer Science 9, (2012) 27-36.
3. García Adeva J.J, Pikatza Atxa J.M , Ubeda Carrillo M, Ansuategi Zengotitabengoa E. Automatic text classification to support systematic reviews in medicine. Expert Systems with Applications, Elsevier 41 (2014)1498–1508.
4. Jiali Yun, Liping Jing , Jian Yu, Houkuan Huang, A multi-layer text classification framework based on two-level representation model. Expert Systems with Applications 39 (2012) 2035–2046.
5. Gary Doran · Soumya Ray, "A theoretical and empirical analysis of support vector machine methods for multiple-instance classification", Mach Learn 97: Springer (2014)79–102.
6. Selcuk Korkmaz, Gokmen Zararsiz, Dincer Goksuluk**,** Drug/nondrug classification using Support Vector Machines with various feature selection strategies**,** computer methods and programs in biomedicine (2014).
7. Subhransu Maji, Alexander C. Berg, and Jitendra Malik, Efficient Classification for Additive Kernel SVMs, IEEE transactions on pattern analysis and machine intelligence (2012).
8. Jingnian Chen, Caiming Zhang, Xiaoping Xue, Cheng-Lin Liu, Fast instance selection for speeding up support vector machines, Knowledge-Based Systems (2013)1–7.
9. Chih-FongTsai , Che-WeiChang. SVOIS: Support Vector Oriented Instance Selection for text classification, Information Systems 38 (2013)1070–1083.
10. B.Ramesh, J.G.R.Sathiaseelan, Support Vector Machine using Efficient Instance Selection for Micro Array Datasets, 2014 IEEE Conference, (2014)644-647.
11. Xinjun Peng, Dong Xu, Structural regularized projection twin support vector machine for data classification, Information science, Elesever (2014).

12. Qiaolin Ye, Ning Ye and Tongming Yin, Enhanced multi weight vector projection support vector machine, Pattern Recognition Letters 42, (2014)91-100.
13. Chin Heng Wan, Lam Hong Lee, Rajprasad Rajkumar, Dino Isa, A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine, Expert Systems with Applications 39 (2012)11880–11888.
14. Marcin Czajkowskia, Marek Grzesb, Marek Kretowskiaa. "Multi-test decision tree and its application to micro array data classification". Artificial Intelligence in Medicine 61 (2014)35–44.
15. Chih-Fong Tsai, Zong-Yao Chen, Shih-Wen Ke, Evolutionary instance selection for text classification, The Journal of Systems and Software 90 (2014)104–113.